Sampling Informative Positives Pairs in Contrastive Learning

Melanie Weber
*Harvard University**
Cambridge, MA, United States
mweber@seas.harvard.edu

Philip Bachman
Montreal, Canada
phil.bachman@gmail.com

***Abstract*—Contrastive Learning is a paradigm for learning representation functions that recover useful similarity structure in a dataset based on samples of positive (similar) and negative (dissimilar) instances. The quality of the learned representations depends crucially on the degree to which the strategies for sampling positive and negative instances reflect useful structure in the data. Typically, positive instances are sampled by randomly perturbing an anchor point using some form of data augmentation. However, not all randomly sampled positive instances are equally effective. In this paper, we analyze strategies for sampling more effective positive instances. We consider a setting where class structure in the observed data derives from analogous structure in an unobserved latent space. We propose *active sampling* approaches for positive instances and investigate their role in effectively learning representation functions which recover the class structure in the underlying latent space.**

***Index Terms*—Contrastive Learning, Active Learning**

I. Introduction

Representation learning is a central task in Machine Learning which seeks to identify useful structure in high-dimensional data and encode this structure in representations that are useful for downstream tasks. Representations may be learned from labeled (supervised learning) or unlabeled training data (unsupervised learning). Recently, *contrastive learning* has received a surge of interest in the representation learning community [van den Oord et al., 2018, Bachman et al., 2019, Chen et al., 2020, Khosla et al., 2020, Zbontar et al., 2021]. In contrastive learning, a representation function is trained to approximately recover similarity structure in the data using pairs of similar and dissimilar observations as a proxy. Applications in a wide range of domains [van den Oord et al., 2018], including Computer Vision [Bachman et al., 2019, Ma et al., 2021], Natural Language Processing [Gao et al., 2021, Meng et al., 2021] and Medical Imaging [Fedorov et al., 2021], have demonstrated the promise of contrastive learning.

The representation function is trained using samples of positive (similar) and negative (dissimilar) instances. Positive instances may be generated via data augmentations such as cropping, flipping or blurring of images (Computer Vision) or via sampling of adjacent sentences (Natural Language Processing). Negative instances are often generated by randomly sampling other images in a data set or other sentences in a corpus. The quality of the resulting representation function depends crucially on the quality of the positive and negative instances used for training, i.e., how informative they are about salient structure in the underlying distribution. In this paper, we investigate how the choice of sampling strategies may impact the quality of the learned representation function and how we can use such insight to design effective sampling strategies. We focus on the sampling of *positive* instances. We consider a setting where the learner has access to high-dimensional observations generated from unobserved latent variables that are sampled from a hidden latent space. We seek to train a representation function which approximately recovers low-dimensional structure (latent classes) in the hidden latent space (Figure 1). In practise, representation functions are evaluated with respect to how well simple prediction functions built on top of them perform on downstream tasks. Here, we analyze the quality of our learned representation function with respect to its ability to recover class structure in the hidden latent space. Our contributions are as follows:

1) First, we propose and analyze an iterative contrastive learning approach (Alg. 1) which seeks to select positive instances that uncover weaknesses in the representation function. We propose to actively sample instances near the decision boundary of a classifier trained in the reconstructed latent space. We present both theoretical and experimental evidence for the strategy's effectiveness.
2) Second, we assume that useful structure in the observations depends strongly on an *informative* subset of latents and weakly on the remaining *uninformative* latents. We study active sampling strategies (Alg. 2) that generate positive instances with a bias towards capturing variability in the informative subset of latents. We present experimental evidence for the effectiveness of such a strategy.

II. Background and Notation

A. Latent Variable Model

We consider the following latent variable model (illustrated in Fig. 1): Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the space of observations and $\mathcal{Z} \subseteq \mathbb{R}^k$ ($k \leq d$) the space of latent variables. We observe $x \in \mathcal{X}$, generated by a map $g : \mathcal{Z} \to \mathcal{X} \in \mathcal{G}$, which we assume to be smooth and nonzero. We follow standard convention in the literature and identify $\mathcal{Z}$ with the unit hypersphere $\mathbb{S}^{k-1}$. Recall that $\mathbb{S}^{k-1} = \{z \in \mathbb{R}^k : z^T z = 1\}$. The metric on

*Work done while intern at Microsoft Research Montreal.

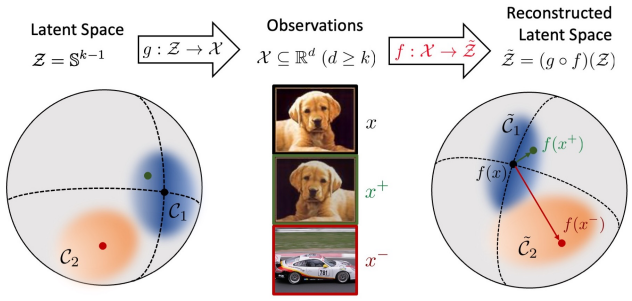

Fig. 1: **Contrastive learning in the latent variable model.**

$\mathbb{S}^{k-1}$ is given by the great-circle distance, i.e., $d(z, z') = \arccos(z^T z') \in [0, \pi] \quad \forall z, z' \in \mathbb{S}^{k-1}$.

In the following we assume a latent class structure in $\mathcal{Z}$, i.e., we have classes $\mathcal{C} = \{C_1, C_2, \dots\} \subseteq \mathcal{Z}$ (class labels are denoted with lower case letters, i.e., $c_1, c_2, \dots$.), each defined by a certain region on the hypersphere, which, geometrically, corresponds to a spherical cap. Let $c \sim p_c$ denote a distribution over the latent classes and $z \sim p_c(z)$ the uniform distribution over the class c. Furthermore, $(z, z^+) \sim p(z^+|z)p_c(z)$ denotes a positive pair of latent variables sampled conditioned on the class c, where we first sample an *anchor* point z given c and then a positive instance z^+ conditioned on z. Negative instances are sampled from the marginal over $\mathcal{Z}$, i.e., by drawing $c' \sim p_c$, $z^- \sim p_{c'}(z)$. Let $p_{data}(\cdot|c)$ denote a data distribution on $\mathcal{X}$ and $p_{pos}(\cdot|x, c)$ a distribution of positive instances x^+ on $\mathcal{X}$, conditioned on an anchor point x. For any observation $x \in \mathcal{X}$, $p(z|x, c)$ denotes the (unknown) posterior distribution over the true latent variables that generated x.

B. Contrastive Loss

Our goal is to learn a representation function $f : \mathcal{X} \to \mathbb{S}^{k-1} \in \mathcal{F}$ that recovers the latent variables from observations. In particular, f defines a distribution $p(\tilde{z}|x, c)$ over reconstructed latent variables given an observation x. We want to learn f contrastively, i.e., by minimizing the *contrastive loss*:

$$\mathcal{L}_{contr}(f; \tau, m) := \tag{II.1}$$
$$\mathop{\mathbb{E}}_{\substack{(x,x^+)\sim p_{pos} \\ x_i^- \sim p_{data}}} \left[-\log \frac{e^{\frac{f(x)^T f(x^+)}{\tau}}}{e^{\frac{f(x)^T f(x^+)}{\tau}} + \sum_{i=1}^m e^{\frac{f(x_i^-)^T f(x^+)}{\tau}}} \right].$$

The contrastive loss (II.1) learns representations that are similar for positive pairs and dissimilar for randomly sampled negative pairs. In the following, we denote the reconstructed latent space as $\tilde{\mathcal{Z}}$ and use the notation $\tilde{z} := h(z) = (f \circ g)$, where $f \approx \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{L}_{contr}(f)$. The superscript $\sim$ will indicate an object in the reconstructed latent space. The minimum $f^* \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{L}_{contr}(f)$ can be seen as an inversion of the generative map g [Zimmermann et al., 2021], i.e., the composition $h^* := f^* \circ g$ preserves the alignment between latent variables. Note that this requires that h^* preserves the dot products between positive pairs (z, z^+) up to a constant, i.e., $\kappa z^T z^+ = h^*(z)^T h^*(z^+)$ (with $\kappa > 0$), which is equivalent to requiring that h^* locally reconstructs the latent space up to linear and orthogonal transformation. Further details are given in the supplemental (sec. S.1).

III. Informative positive instances for recovering latent class structure

Our goal is to identify a class structure $\tilde{\mathcal{C}} = \{\tilde{C}_1, \tilde{C}_2, \dots\} \subseteq \tilde{\mathcal{Z}}$ in the reconstructed latent space that recovers the latent classes $\mathcal{C} = \{C_1, C_2, \dots\} \subseteq \mathcal{Z}$ in the true latent space. We propose contrastive learning approaches that sample positive instances, which are *informative* about the latent class structure, with the goal of training representation functions more efficiently. We first consider a setting where we know the number of classes $\mathcal{C}$ and that they are linearly separable. Our iterative contrastive learning algorithm (Alg. 1) samples m_t new positive and negative instances in round t and trains an updated representation function f_t using initialization f_{t-1}. This approach will compute some approximation $\hat{f} \approx \min_{f \in \mathcal{F}} \mathcal{L}_{contr}$ to the global minimizer f^*. We define the reconstruction error of $\hat{f}$ with respect to erroneous class labels: we say that $\hat{f} \in \mathcal{F}$ is α-*accurate*, if for $z \sim p_z$ we have that $z \in C_i$ implies $\tilde{z} = (g \circ f)(z) \in \tilde{C}_i$ with probability $1 - \alpha$.

The key idea of our active sampling approach is to sample positive instances that uncover the weaknesses of $\hat{f}$. Fig. 2 illustrates the motivation behind our approach schematically: Due to the inaccuracies in the representation function, it may happen that positive pairs are correctly labeled as being in the same class in the true latent space, but not in the reconstructed latent space. For example, instances $x^+ \in \mathcal{X}$ generated from latents that are close to a hyperplane separating two classes in $\mathcal{Z}$ (marked in red) could help us identify inaccuracies in $\hat{f}$, if their representations $\hat{f}(x^+)$ are misclassified by a classifier learned in $\tilde{\mathcal{Z}}$. We formalize this observation below and describe sampling approaches for generating such "informative" positive instances. Below, we describe three strategies for sampling positive instances (Alg. 1(i)-(iii) and Fig. 2).

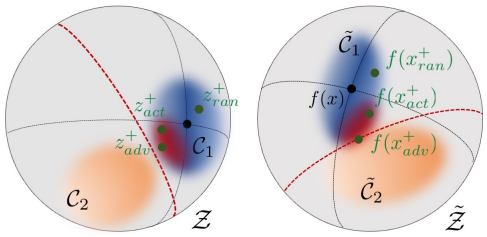

Fig. 2: **Sampling positive instances.** Positive instances for an anchor point (black), sampled actively (x^+_{act}) or adversarially (x^+_{adv}), can improve over random augmentation (x^+_{ran}). The image of actively instances ($f(x^+_{act})$) lies close to the decision boundary of our classifier q_t (acceptance region shown in red). The image of adversarial instances ($f(x^+_{adv})$) is closest to the decision boundary or may transgress it if the reconstructed classes are not separable due to representation error.

A. Sampling strategies for positive pairs

1) Random augmentation: In practise, positive instances are often sampled via random augmentation of an anchor point. The augmentation method is highly dependent on the application domain: For images, the anchor might be rotated or cropped to generate a positive instance. In video analysis,

Algorithm 1 Active selection of positive pairs

1: Learn initial $f_0 \in \mathcal{F}$ and classifier q_0 based on $\mathcal{D}_0$.
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: **for** $i = 0, \ldots, m_t - 1$ **do**
4: Sample anchor $c \sim p_c$, $x_i \sim p_{data}(\cdot|c)$.
5: (i) **Random:** Sample $x_i^+ \sim p_{pos}(\cdot|x_i, c)$.
6: (ii) **Active:** Sample $x_i^+ \sim p_{pos}(\cdot|x_i, c)$.
7: If $\left|w_t \cdot f(x_i^+)\right| \geq a_t$: **Continue**.
8: (iii) **Adversarial:**
9: Set $x_i^+ \leftarrow \text{argmax}_{x \in \mathcal{X}}\, l\left(\{q_t(f_t(x)) - q_t(f_t(x_i))\}\right)$.
10: **end for**
11: $\mathcal{D}_{t+1} := \{(x_i, x_i^+; x_i^- \sim p_{data})\}_{j=1}^{m_t} \cup \mathcal{D}_t$
12: Learn $f_{t+1} \in \text{argmin}_{f \in \mathcal{F}} \mathcal{L}_{contr}(f)$ over $\mathcal{D}_{t+1}$, with initialization f_t. Learn q_{t+1}.
13: **end for**

one may choose adjacent frames as positive pairs. Similarly, when working with text, positive pairs may be generated by choosing adjacent sentences or dropping different subsets of words from the same text sequence. Our baseline approach emulates these classical augmentation techniques in an abstract setting: we begin by sampling an anchor point x, generated from a randomly selected latent class. The positive instance is then generated via augmentation, in our case random perturbation of the anchor point. We evaluate the quality of our representation function by testing its ability to recover the latent class structure. For this, we learn a classifier $q : \tilde{\mathcal{Z}} \to \mathbb{R}^{|\mathcal{C}|}$, $q(f(x)) := Wf(x) = W(f \circ g)(z)$, in the reconstructed latent space. Our heuristic description of the class structure at the beginning of the section can be formalized as follows: Let $\gamma > 0$ denote a margin, such that there exist $w_1, \ldots, w_{|\mathcal{C}|} \in \mathbb{R}^d$ with $\sum_{i=1}^{|\mathcal{C}|} \|w_i\|^2 < 1$, such that for latents $z \sim p_c$, we have $\langle z, w_c \rangle \geq \frac{\gamma}{2}$ and $\langle z, w_{c'} \rangle \leq -\frac{\gamma}{2}$ for all $c \neq c'$, i.e., the latent classes $\mathcal{C}$ are separable. Given access to observations $x \in \mathcal{X}$, generated by g, and a contrastively learned representation function f that reconstructs the latent space $\tilde{\mathcal{Z}}$, we can learn an "optimal" matrix $W \in \mathbb{R}^{|\mathcal{C}| \times d}$ to obtain the classifier q in $\tilde{\mathcal{Z}}$. In this setting, we have the following generalization bound on the performance of the contrastively learned representations:

Theorem 3.1: Let $\gamma > 0$ (fixed) denote the margin in the true latent space and $\tilde{\gamma} \leq \gamma$ the margin in the reconstructed latent space. For any $\delta > 0$ we have with probability at least $1 - \delta$ that ($\rho' := 1/(1-\rho)$)

$$L_{class}(f) \leq \rho' \left(\hat{\mathcal{L}}_{un}(f) - \rho\right) + \rho' \left(4 L_\alpha \mathcal{R}_{\mathcal{D}}(\mathcal{F}) + \sqrt{\frac{\log \frac{1}{\delta}}{2|\mathcal{D}|}}\right)$$

for all $f \in \mathcal{F}$. Here, L_{class} characterizes the classification error, $\hat{\mathcal{L}}_{un}$ the empirical error; $\mathcal{R}_{\mathcal{D}}(\mathcal{F})$ denotes the Rademacher complexity of the function class $\mathcal{F}$ and $L_\alpha \leq \frac{2}{\tilde{\gamma}}(1 - 2\alpha)$ and ρ the probability of sampling a false negative instance, i.e., the probability of sampling twice from the same class.

The theorem builds on an extension of [Saunshi et al., 2019, Thm. 4.1] to our latent variable model described above. We defer all proof details to the supplemental (sec. S.2.2).

Positive instances (x, x^+) sampled via random augmentation will differ in their effectiveness, with some being more informative about the underlying latent class structure than others: consider a set $\mathcal{D}_0 \subseteq \mathcal{Z}$, which contains samples from two latent classes that are linearly separable with margin γ. If our contrastive approach learns a representation function f_t that is α-accurate, then, due to representation error, $h(\mathcal{D}_0) \subseteq \tilde{\mathcal{Z}}$ will be separable with a smaller margin $\tilde{\gamma} < \gamma$ only. Hence, instances that lie close to the decision boundary will be more challenging to classify in that they preserve the class label in the true latent space $\mathcal{Z}$, but not in the reconstructed latent space $\tilde{\mathcal{Z}}$. Such positive pairs are particularly informative for learning good representations. Can we encourage the sampling of such positive instances over less informative ones?

2) Active selection of positive instances: The first alternative approach that we discuss selects positive instances *actively*. In round t we can learn a classifier q_{t-1} in $\tilde{\mathcal{Z}}$, which is consistent with the representations of instances in $\mathcal{D}_{t-1}$. Then we sample a new anchor point x and generate a positive instance x^+ via random augmentation. In contrast to the baseline approach, we accept the positive instance only if it is close to the decision boundary of q_{t-1}. Instances that are far away from the decision boundary are assumed to be less informative and will be rejected (Alg. 1, line 6). This setting encourages the sampling of positive instances of the type described above. Recall that, in practise, we do not have access to the error of the current representation function (i.e., to the parameter α) and consequently also not to the margin $\tilde{\gamma}$ in the reconstructed latent space. However, we know that the desired positive instances lie near the decision boundary of our classifier q_t, so we can encourage our algorithm to sample from this region (Fig. 2, marked red; characterized by choice of hyperparameter a_t in Alg. 1). Below, we will see experimental evidence for the advantage of the active strategy over the baseline (see Table I). We can also analyze the advantage of active sampling theoretically. For this, we pick two classes and focus on the problem of learning a separator between them. This reduces the problem to a binary classification task. (An extension to the multi-class case follows from analogous arguments.) Formally, let $|\mathcal{C}| = 2$ and $C_\pm := \{z | (z, \pm 1) \in \mathbb{S}^{k-1} \times \mathcal{C}\} \subseteq \mathcal{Z}$. We assume that C_+, C_- are linearly separable with margin γ. Our goal is to to learn a classifier in $\tilde{\mathcal{Z}}$ that recovers the latent structure defined by $\{C_+, C_-\} \subseteq \mathcal{Z}$, i.e., we want to learn a classifier that separates $\{\tilde{C}_+, \tilde{C}_-\} \subseteq \tilde{\mathcal{Z}}$. Note that our active sampling strategy for positive instances resembles classical active learning techniques for binary classification [Balcan et al., 2007]. To compare the efficiency of the active sampling strategy and the baseline, we compare the number of samples (m_t) that we need to draw in each round to ensure convergence. We see that the active approach allows for a much reduced sample size, indicating that representation functions can be trained more efficiently via active selection of positive instances. Formally, we have the following result:

Theorem 3.2: For any $\delta, \epsilon > 0$, we can recover the class structure up to error ϵ with probability $1 - \delta$ with (1) sample complexity $m = O(\frac{d}{\epsilon})$ for random augmentations (Alg. 1(i)) and (2) sample complexity $m = O\left(d^{3/2} \log(\frac{1}{\epsilon})\right)$ for active selection (Alg. 1(ii), with rejection threshold $a_t = \frac{\pi}{2^{t-1}}$).

The parameter a_t may be chosen via hyperparameter search.

We defer all proof details to the supplemental (sec. S.2.3).

3) Adversarial augmentation: We briefly discuss a second active sampling strategy, which generates positive instances via *adversarial augmentations*. Here, we seek to select the most margin-transgressing positive instance by applying targeted perturbations to the anchor. In round t, we locally maximize the loss with respect to the classifier q_t that we learned based on the representation function f_t (Alg. 1(iii)), i.e., we solve

$$\operatorname*{argmax}_{x \in \mathcal{X}} \; l\left(\{q_t(f_t(x)) - q_t(f_t(x_i))\}\right) \tag{III.1}$$

with a suitable loss function l (e.g., squared loss). The selection of such targeted augmentations may further reduce the number of labels m that we need to add in each iteration. However, solving the optimization problem in Eq. III.1 for each positive instance iterations is computationally challenging in practise.

B. Experiments

Parameters			Baseline	Active	Double-active
n_{lat}	n_{obs}	N			
48	256	200	0.937	**0.951**	0.945
48	512	200	0.946	**0.955**	**0.955**
64	512	200	0.907	**0.916**	0.901
64	256	500	0.903	0.907	**0.927**
64	1024	500	0.922	0.920	**0.927**

TABLE I: Results for Algorithm 1 for two latent classes, n_{lat} latent dimensions, n_{obs} observed dimensions. We report the highest classification accuracy reached after N updates.

We train representation functions with positives sampled in the reconstructed latent space. In each iteration t, positive and negative instances are sampled for a set of anchor points and then added to the training data $\mathcal{D}_t$. We then train a representation function f_t using the standard contrastive loss (Eq. II.1). Representations are evaluated in terms of their ability to recover the latent class structure. For this, we learn a classifier q_t in the reconstructed latent space and report its accuracy. All data is synthetic and sampled from the unit hypersphere. To define class-conditioned distributions in the true latent space, we define a class mean on the unit hypersphere, add isotropic Gaussian noise with a predefined standard deviation and renormalize to unit length. We investigate the following four scenarios, which employ different passive and active sampling strategies:

1) **Baseline:** Anchor points and positive instances are sampled from the baseline prior (resembling Alg. 1(i)).
2) **Active:** Anchor points are sampled from the baseline. Candidate positive instances are sampled via perturbation in the observation space until one is found whose image lies in the acceptance region, i.e., within some ϵ of the decision boundary of q_{t-1} in the reconstructed latent space. This resembles the *active selection* strategy (Alg. 1(ii)).
3) **Double-active:** Anchor points sampled with a bias towards the region near the decision boundary of q_{t-1} in the reconstructed latent space. Candidate positive instances are sampled via perturbation in the observation space until one is found whose image lies in the acceptance region. With the additional preference for sampling near the decision boundary, this can be seen as closer to the *adversarial augmentation* idea (Alg. 1(iii)).

The results are given in Table I. We see that the active strategies consistently outperform the baseline. This corroborates the theoretical results (Thm. A.7), which shows that active strategies can converge at a faster rate.

IV. Sampling positive instances via targeted augmentations in latent space

Algorithm 2 Active perturbation in latent space

Initialize $f_0 \in \mathcal{F}$.
Define index set $I \subset [k]$ of informative latent variables.
for $t = 0, 1, \ldots, T-1$ **do**
  for $i = 0, \ldots, m-1$ **do**
    Sample anchor $c \sim p_c$ and $z, \bar{z} \sim p_{data}(\cdot|c)$.
    Calibrate τ.
    Perturb informative latents: $z_I^+ \leftarrow p_\tau(\cdot|z_I)$
    Randomize other latents: $z_{\neg I}^+ \leftarrow \bar{z}_I$
    $(z_i, z_i^+) \leftarrow (z, z^+)$
  end for
  $\mathcal{D}_t := \{(x_i, x_i^+; x_i^- \sim p_{data})\}_{j=1}^m$
   with $x_i = g(z_i), x_i^+ = g(z_i^+)$.
  Learn $f^{t+1} \in \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{L}_{contr}(f)$, with initialization f^t.
end for

In this section we consider a second setting for sampling informative positive instances. Contrary to the previous setting, we now study augmentations in the true latent space rather than the observation space or the reconstructed latent space. For example, imagine an agent interacting with a visual scene where some of the latent variables describe useful structure in the scene's appearance (e.g., type and location of various objects), while other latent variables describe more superficial structure (e.g., surface texture and shading of each object). Here, we may want to learn representations, which focus on structure in the former latent variables rather than the latter.

A. Sampling strategies

In this setting, interventions on the data generation process (e.g., when the agent interacts with an object in the scene) are analogous to data augmentations in our earlier settings, where the values of some latent variables are fixed and others are perturbed. In practise, useful structure in the observations does not depend equally on all latent variables and is largely determined by a particularly *informative* subset. In such cases, it can be helpful to sample positive instances via perturbations that focus informative latents, rather than perturbations, which affect the latents more uniformly. Scenarios with such an informative subset of latents may arise in many Reinforcement Learning settings: Suppose an agent receives complex observations, which are functions of a large number of latents, but can only influence a subset of them through its actions. In this case, characterizing variability in the controllable subset is critical for choosing optimal actions, while the other latents evolve independently of the agent's actions and are thus less important

for planning. In our framework, the latents influenced by the agent's actions are *informative* with respect to solving a control task, whereas the remaining latents are *uninformative*. While we may not know precisely what latent variables there are, or which latent variables are informative, we know that the agent can only manipulate the informative latents. Hence, we may assume that the agent can perform targeted perturbations of informative latents. Formally, we investigate sampling strategies for positive pairs $p(z^+|z)$ (with anchor $z \sim p(z)$). Their effectiveness is again evaluated with respect to the encoder's ability to learn a posterior distribution $p(\tilde{z}|x,c) =: q_h(z|x,c)$, such that structure in the true latent space (encoded in the unknown distribution $p(z|x,c)$) is recovered in the reconstructed latent space:

Theorem 4.1: Consider the minimizer $h^* = \operatorname{argmin}_{(z,z^+)} \mathbb{E}\left[H(p(z^+|z,c), q_h(z^+|z,c))\right]$ *with* $q_h(z^+|z,c) = C_h(z^+)^{-1} e^{h(z^+)^T h(z)/\tau}$ *denoting the conditional distributions (where* $C_h(z) := \int e^{h(z^+)^T h(z)/\tau} dz$*) over reconstructed latent variables and* $H(p, q_h)$ *denoting the cross entropy between distributions* p *and* q_h*. Then* h^* *locally reconstructs latent space up to linear and orthogonal transformation.*

The proof adapts a result by Zimmermann et al. [2021]; we defer all details to supplemental S.1. In the following, let the informative subset of latent variables be characterized by an index set $I \subset [k]$, where k is the dimension of the latent space. Algorithm 2 proposes a sampling strategy where targeted perturbations are applied to the informative latents, whereas other latents are randomized. We will give experimental evidence for such an *active perturbation* strategy in the following section.

Parameters			Baseline	Info-active	Active	Class-preserve
n_c	n_i	n_n				
4	4	4	0.777	**0.956**	0.936	0.986
8	4	4	0.609	**0.867**	0.817	0.900
4	2	6	0.728	**0.967**	0.942	0.974
8	2	6	0.408	**0.692**	0.611	0.687
4	2	16	0.759	**0.969**	0.949	0.975

TABLE II: Results for Alg. 2 with n_c classes, n_i informative latents, n_n noisy latents. We report kNN accuracy averaged over 10 resamples of the train and test sets after 1000 updates.

B. Experiments

Our experimental setup trains a representation function with positive pairs sampled according to Algorithm 2. We define class-conditioned distributions in the true latent space by concatenating samples from the surface of two unit hyperspheres, one of dimension n_{info}, representing the informative latents, and one of dimension n_{noise}, representing the uninformative latents. For sampling informative latents, we define a class mean on the first hypersphere, add isotropic Gaussian noise with a predefined standard deviation and then renormalize to unit length. With that, the distribution of informative latents conditioned on a class resembles a von Mises-Fisher distribution (see Eq.(1.1.) in the supplemental, sec. 1). The standard deviation is calibrated to generate little overlap between the classes in order to simulate separability. The uninformative latents are sampled uniformly at random from the second hypersphere. The scale of the uninformative latents can be varied to control the degree to which variability in the observation space is due to the informative or uninformative latents. The effect of the latents on the observations is entangled and non-linear due to the function $g(z)$, so identifying informative vs uninformative variability via observations $x = g(z)$ is non-trivial. We investigate different scenarios for sampling positive instances given an anchor:

1) **Baseline:** Equal-sized random perturbations of all latents.
2) **Info-active:** Small, targeted perturbations of informative latents and independent random sampling of all others.
3) **Active:** Small, targeted perturbations of informative latents and larger, targeted perturbations of all others.
4) **Class-preserving:** Assume access to a class-preserving transformation, which is used to augment the anchor to generate a positive sample from the same class. This is an idealized setting, which can be thought of as a supervised approach. It is included for comparison; unsupervised approaches are not expected to match its accuracy.

Our results confirm that targeted perturbation of the informative latent variables generate more effective positive instances than the baseline (Tab. II, baseline vs. info-active). We further see that even a small difference in the perturbation scale of informative and uninformative latents can improve over the baseline (Tab. II, baseline vs. active). Details and further results are deferred to supplemental S.3. A strong advantage of this active sampling strategy over the methods discussed in previous section is that they *does not require prior knowledge of classes or class structure*.

V. Conclusions

In this paper, we studied contrastive learning approaches for recovering low-dimensional structure in a latent variable model. We proposed and analyzed different sampling approaches for generating effective positive instances, both theoretically and in exploratory experiments. Our results confirm the intuition that encouraging the sampling of positive pairs that are *informative* with respect to the underlying class structure or that explicitly uncover weaknesses in the representation function can improve the efficiency of the contrastive learning approach.

We investigate the sampling of positive instances in an abstract setting modeled after popular sampling strategies used in practise. In particular, our *random augmentation* approach is an abstraction of sampling strategies used in image and video analysis, where images or video frames are randomly cropped or blurred [Chen et al., 2020]. An interesting avenue for future investigation is the adaption of the proposed sampling techniques to application domains (Computer Vision, Natural Language Processing), which would allow to evaluate their efficiency in a more practical setting. We further investigate a setting where useful structure in the observations is largely driven by a subset of latent variables (*informative latents*). This setting is motivated by Reinforcement Learning and control tasks, where an agent's actions only influence a subset of the latents. Testing our proposed info-active sampling strategy in one of these settings is an interesting direction for future work.

References

M. Anthony, L.M.M. Anthony, P.L. Bartlett, P.L. Bartlett, and Cambridge University Press. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Dipendra Misra. Investigating the role of negatives in contrastive representation learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 7187–7209, 28–30 Mar 2022.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems 32*, pages 15535–15545. 2019.

Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In *International Conference on Computational Learning Theory*, pages 35–50. Springer, 2007.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021.

Alex Fedorov, Lei Wu, Tristan Sylvain, Margaux Luck, Thomas P. DeRamus, Dmitry Bleklov, Sergey M. Plis, and Vince D. Calhoun. On self-supervised multimodal representation learning: An application to alzheimer's disease. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1548–1552, 2021.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv:2104.08821*, 2021.

Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised constrastive learning. In *International Conference on Machine Learning*, pages 3821–3830. PMLR, 2021.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

Tyler L Hayes, Maximilian Nickel, Christopher Kanan, Ludovic Denoyer, and Arthur Szlam. Can i see an example? active learning the long tail of attributes and relations. *arXiv preprint arXiv:2203.06215*, 2022.

Chih-Hui Ho and Nuno Nvasconcelos. Contrastive learning with adversarial examples. *Advances in Neural Information Processing Systems*, 33:17081–17093, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. In *International Conference on Learning Representations*, 2021.

Yu Meng, Chenyan Xiong, Payal Bajaj, saurabh tiwary, Paul N. Bennett, Jiawei Han, and Xia Song. COCO-LM: Correcting and contrasting text sequences for language model pretraining. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, page 1196–1204, 2013.

Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. On compositions of transformations in contrastive self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9577–9587, 2021.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021.

Li S. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics & Statistics*, 4, 01 2011.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5628–5637. PMLR, 09–15 Jun 2019.

Anshul Shah, Suvrit Sra, Rama Chellappa, and Anoop Cherian. Max-margin contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8220–8230, 2022.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pages 10268–10278. PMLR, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018.

Feng Wang, Huaping Liu, Di Guo, and Sun Fuchun. Unsupervised representation learning by invariance propagation.

In *Advances in Neural Information Processing Systems*, volume 33, pages 3510–3520, 2020.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. *arXiv:2203.13457*, 2022.

Seunghan Yang, Debasmit Das, Simyung Chang, Sungrack Yun, and Fatih Porikli. Distribution estimation to automate transformation policies for self-supervision. *arXiv:2111.12265*, 2021.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly supervised contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10042–10051, October 2021.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021.

APPENDIX

A. Related Work

Contrastive learning has received a surge of interest in the last few years. A large body of work investigates contrastive learning methods empirically [van den Oord et al., 2018, Bachman et al., 2019, Dwibedi et al., 2021, Chen et al., 2020]. Many recent works focus on understanding the impact of different approaches for sampling positive [Ho and Nvasconcelos, 2020, Tian et al., 2021, Zheng et al., 2021, Hayes et al., 2022, Wang et al., 2020] and negative instances [Chuang et al., 2020, Ma et al., 2021, Ash et al., 2020, 2022, Shah et al., 2022, Robinson et al., 2021], motivated by the intuitive idea that sampling instances that are more *informative* about the underlying structure could lead to more effective contrastive learning. Notably, Ho and Nvasconcelos [2020] propose to generate "challenging" positive pairs via adversarial perturbation (*adversarial augumentation*), which they demonstrate empirically to be promising. Zheng et al. [2021] empirically investigate an approach that selects positive instances via a graph-based weakly-supervised approach. To the best of our knowledge, sampling approaches for positive instances have not been studied systematically in a latent variable model. Yang et al. [2021] and Patrick et al. [2021] investigate approaches for optimizing the composition of transformations.

The theoretical analysis of contrastive learning approaches has recently received increasing attention [Saunshi et al., 2019, HaoChen et al., 2021, Graf et al., 2021, Wang and Isola, 2020, Zimmermann et al., 2021, Wang et al., 2022]. Notably, [Saunshi et al., 2019] proposed one of the first theoretical frameworks for contrastive learning, which evaluates the quality of the learned representations on a downstream classification task. However, they make no assumptions on the structure of the underlying latent space. Wang and Isola [2020] and Zimmermann et al. [2021] analyzed contrastive learning in a latent variable model, albeit without assuming additional structure, such as latent classes. Both works consider only classical sampling strategies, where positive instances are generated via random augmentations.

B. Reconstruction of Latent Space

We first introduce a framework for analyzing the quality of a representation function with respect to its ability to recover the latent class structure in $\mathcal{Z}$. Recall that we identify the latent space with the unit hypersphere, i.e., $\mathcal{Z} = \mathbb{S}^{k-1}$. The latent classes $\mathcal{C} = \{C_1, C_2, \dots\}$ form spherical caps in $\mathcal{Z}$ (see Fig. 1, main text). We denote class labels with lower case letters, i.e., $c_1, c_2, \dots$. In the following, we assume that the conditional distribution of positive pairs of latent variables (z, z^+) from which (x, x^+) are generated is von Mises-Fisher, i.e.,

$$p(z^+|z) = C_p^{-1} \exp(\tau^{-1} z^T z^+) \tag{A.1}$$

$$C_p = \int \exp(\tau^{-1} z^T z^+) dz^+ \,, \tag{A.2}$$

where $(z, z^+) \sim p(z^+|z)p_c(z)$ is a positive pair of latent variables sampled from class c and $\tau > 0$ a hyperparameter. The marginal distribution over the class c is assumed to be uniform, i.e., $p_c(z) = |\mathcal{C}|^{-1}$. Recall that observations $x \in \mathcal{X}$ are generated by an unknown map g, i.e., $x = g(z)$. We can define a posterior distribution $p(z|x)$ over the true latent variables that generated x. In Algorithms 1(i) and 2, positive instances are directly sampled from $p(z^+|z)$. In Algorithms 1(ii) and 1(iii), candidate positive instances are drawn from $p(z^+|z)$ and accepted according to the specified rules. Negative instances are sampled uniformly at random.

Recall that we want to learn a representation function $f : \mathcal{X} \to \mathbb{S}^{k-1}$, such that the composition $h : \mathbb{S}^{k-1} \to \mathbb{S}^{k-1}$, $h = f \circ g$ preserves the alignment between latent variables. A good representation function recovers the hidden latent variables, the underlying task is a demixing problem, where we learn to invert the generative process g (up to orthogonal linear transformations). This requires that h preserves the dot products between positive pairs (z, z^+) up to a constant, i.e., $\kappa z^T z^+ = h(z)^T h(z^+)$ (with $\kappa > 0$). This is equivalent to requiring that h locally reconstructs the latent space up to linear and orthogonal transformation.

In the absence of class structure, [Zimmermann et al., 2021, Prop.1] showed that if $\mathcal{F}$ is sufficiently rich, a suitable h minimizes the cross entropy of the ground-truth conditional distribution $p(z^+|z)$ and the conditional distribution of the recovered latent variables. In the presence of class structure, an analogous result can be shown for the distribution of positive pairs sampled from a class c:

Theorem A.1: Consider the minimizer

$$h^* = \operatorname{argmin} \mathop{\mathbb{E}}_{\substack{(z,z^+) \\ \sim p(z^+|z)p_c(z)}} \left[H(p(z^+|z,c), q_h(z^+|z,c)) \right] \,, \tag{A.3}$$

with

$$q_h(z^+|z,c) = C_h(z^+)^{-1} e^{h(z^+)^T h(z)/\tau} \tag{A.4}$$

$$C_h(z) := \int e^{h(z^+)^T h(z)/\tau} dz \,, \tag{A.5}$$

denoting the conditional distributions over reconstructed latent variables and $H(p, q)$ denoting the cross entropy between distributions p and q. Then h^* locally reconstructs latent space up to linear and orthogonal transformation.

We include a proof for completeness.

$$
\begin{aligned}
&\mathbb{E}_{c\sim p_c}\left[\mathbb{E}_{z\sim p_c(z)}\left[H\left(p(\cdot|z,c), q_h(\cdot|z,c)\right)\right]\right] \\
&\quad= \mathbb{E}_{c\sim p_c}\left[\mathbb{E}_{z\sim p_c(z)}\left[\mathbb{E}_{z^+\sim p(z^+|z,c)}\left(-\log q_h(z^+|z,c)\right)\right]\right] \\
&\quad\overset{(1)}{=} \mathbb{E}_{(z^+,z)\sim p(z^+|z)p_c(z)}\left[-\frac{1}{\tau}h(z^+)^T h(z) + \log C_h(z)\right] \\
&\quad= -\frac{1}{\tau}\mathbb{E}_{(z^+,z)\sim p(z^+|z)p_c(z)}\left[h(z^+)^T h(z)\right] + \mathbb{E}_{z\sim p_c(z)}\left[\log C_h(z)\right] \\
&\quad\overset{(2)}{=} -\frac{1}{\tau}\mathbb{E}_{(z^+,z)\sim p(z^+|z)p_c(z)}\left[h(z^+)^T h(z)\right] + \mathbb{E}_{z\sim p_c(z)}\left[\log \int_{z'} e^{h(z')^T h(z)/\tau} dz'\right] \\
&\quad\overset{(3)}{=} -\frac{1}{\tau}\mathbb{E}_{(z^+,z)\sim p(z^+|z)p_c(z)}\left[h(z^+)^T h(z)\right] + \mathbb{E}_{z\sim p_c(z)}\left[\log\left(|\mathcal{C}|\cdot \mathbb{E}_{z'\sim p_c(z)}\left(e^{h(z')^T h(z)/\tau}\right)\right)\right] \\
&\quad= -\frac{1}{\tau}\mathbb{E}_{(z^+,z)\sim p(z^+|z)p_c(z)}\left[h(z^+)^T h(z)\right] + \mathbb{E}_{z\sim p_c(z)}\left[\log \mathbb{E}_{z'\sim p_c(z)}\left(e^{h(z')^T h(z)/\tau}\right)\right] + \log|\mathcal{C}| \\
&\quad\overset{(4)}{=} -\frac{1}{\tau}\mathbb{E}_{(z^+,z)\sim p(z^+|z)p_c(z)}\left[((f\circ g)(z^+))^T (f\circ g)(z)\right] + \mathbb{E}_{z\sim p_c(z)}\left[\log \mathbb{E}_{z'\sim p_c(z)}\left(e^{((f\circ g)(z'))^T (f\circ g)(z)/\tau}\right)\right] + \log|\mathcal{C}|\,,
\end{aligned}
$$

where in (1) we have inserted the definition of q_h and in (2) the definition of the partition function C_h. In (3) we have multiplied by 1 ($|\mathcal{C}|\,|\mathcal{C}|^{-1}$) and approximated the integral by sampling from $p_c(z) = |\mathcal{C}|^{-1}$. In (4), we have inserted $h = f \circ g$.

By expressing functions of latent variables with the corresponding expressions for observables, we get

$$
-\frac{1}{\tau}\mathbb{E}_{(x,x^+)\sim p_{pos}}\left[f(x^+)^T f(x)\right] + \mathbb{E}_{x\sim p_{data}}\left[\log \mathbb{E}_{x^-\sim p_{data}}\left(e^{f(x^-)^T f(x)/\tau}\right)\right] + \log|\mathcal{C}| = \mathcal{L}_{align}(f;\tau) + \mathcal{L}_{uni}(f;\tau) + \log|\mathcal{C}|\,.
$$

Geometrically, the cross-entropy encodes the concepts of *alignment* and *uniformity*, which are characterized by the following loss functions [Wang and Isola, 2020, Zimmermann et al., 2021]:

- *Alignment:* Positive pairs should be mapped to nearby feature representations. This is captured in the loss:

$$
\mathcal{L}_{align}(f;\tau) = -\frac{1}{\tau}\mathop{\mathbb{E}}_{(x,x^+)\sim p_{pos}}\left[f(x^+)^T f(x)\right]\,. \tag{A.6}
$$

- *Uniformity:* Feature vectors should be approximately uniformly distributed on $\mathbb{S}^{k-1}$ to encourage separability. This is encoded in the loss:

$$
\mathcal{L}_{uni}(f;\tau) = \mathop{\mathbb{E}}_{x\sim p_{data}}\left[\log \mathop{\mathbb{E}}_{x'\sim p_{data}}\left(e^{f(x')^T f(x)/\tau}\right)\right]\,. \tag{A.7}
$$

In particular, we can show the following relation:

Corollary A.2:

$$
\mathbb{E}_{z\sim p_c(z)}\left[H\left(p(\cdot|z), q_h(\cdot|z)\right)\right] = \mathcal{L}_{align}(f;\tau) + \mathcal{L}_{uni}(f;\tau) + \log|\mathcal{C}|\,.
$$

Thm. A.1 guarantees that representations learned via Algorithms 1 and 2 recover the latent space *locally*, i.e., recover the relationship between close-by points within the same class. What can we say about their ability to recover *global* structure, such as the relationship between classes?

To answer this question, we analyze which geometric assumptions are implicitly encoded in the contrastive loss $\mathcal{L}_{contr}$ via $\mathcal{L}_{align}$ and $\mathcal{L}_{uni}$. We find that $\mathcal{L}_{contr}$ encourages representations that recover a homogeneous reconstructed latent space, where the latent classes are well-concentrated and uniformly distributed in the latent space:

Thm. A.1 and Corr. A.2 suggest that minimizing $\mathcal{L}_{contr}$ implies a small uniformity loss ($\mathcal{L}_{uni}$) and alignment loss ($\mathcal{L}_{align}$). A more detailed analysis reveals that a small uniformity loss ensures that the angular separation between classes is not too uneven, favouring a distribution close to the uniform distribution in the true latent space. A small alignment loss implies that the angular sizes of the classes are not too large, i.e., that the classes are well concentrated. This can be seen with the following arguments:

1) A small uniformity loss $\mathcal{L}_{uni}(f)$ ensures that the angular separation between classes is not too uneven. In particular, note that

$$\begin{aligned}\mathcal{L}_{uni}(f) &= \mathbb{E}_{\substack{c\sim p_c \\ x\sim p_{data}(\cdot|c)}}\left[\log \mathbb{E}_{\substack{c'\sim p_c \\ x^-\sim p_{data}(\cdot|c')}}\left(e^{f(x)^T f(x^-)}\right)\right] \\ &= \rho\mathbb{E}_{\substack{c\sim p_c \\ (x,x^-)\sim p_{pos}(\cdot|c)}}\left[\log \mathbb{E}\left(e^{f(x)^T f(x^-)}\right)\right] + (1-\rho)\mathbb{E}_{\substack{c,c'\sim p_c \\ x\sim p_{data}(\cdot|c) \\ x^-\sim p_{data}(\cdot|c')}}\left[\log \mathbb{E}\left(e^{f(x)^T f(x^-)}\right)\right] \\ &= \rho\mathbb{E}_{(c;x,x^-)}\left[\log \mathbb{E}\left(e^{\tilde{z}^T \tilde{z}^-}\right)\right] + (1-\rho)\mathbb{E}_{(c,x),(c',x^-)}\left[\log \mathbb{E}\left(e^{\tilde{z}^T \tilde{z}^-}\right)\right] \\ &= \rho\mathbb{E}_{(c,x)}\left[\log e^{f(x)^T \mu_{\tilde{c}}}\right] + (1-\rho)\mathbb{E}_{(c,c')}\left[\log e^{\mu_{\tilde{c}}^T \mu_{\tilde{c}'}}\right] .\end{aligned}$$

Notably, a small uniformity loss ensures that the angular separation is not too large for any two classes, implying distribution close to the uniform distribution in the true latent space.

2) A small alignment loss implies that the angular sizes of the classes are not too large, i.e., that the classes are well concentrated. For this, note that

$$\mathcal{L}_{align}(f) = \mathbb{E}_{\substack{c\sim p_c \\ (x,x^+)\sim p_{pos}(\cdot|c)}}\left[e^{f(x)^T f(x^+)}\right] = \mathbb{E}_{(c;x,x^+)}\left[e^{\tilde{z}^T \tilde{z}^+}\right] = \mathbb{E}_{(c;x)}\left[e^{\tilde{z}^T \mu_{\tilde{c}}}\right] .$$

This suggests that the classical contrastive loss $\mathcal{L}_{contr}$ may not capture heterogenity between classes or low-dimensional structure in latent space well. Such geometric information could be uncovered by sampling instances that are informative about the underlying structure. This observation motivates the design of active or adversarial sampling strategies that pick informative positive pairs, with the hope of incorporating more geometric information into the training process (Algorithms 1 and 2).

C. Recovering latent class structure

In this section we give theoretical evidence for the quality of the representation functions trained with Algorithm 1. Specifically, we analyze how well the representations recover the underlying latent class structure. We focus on the comparison of passive an active sampling strategies, i.e., *random augmentation* (Algorithm 1(i)) and *active selection* (Algorithm 1(ii)). Both approaches sample positive instances via random augmentation. However, while the random augmentation approach adds each of the sampled instances to the training data, the active selection approach rejects instances that are not close to the decision boundary and therefore less informative. We provide a theoretical argument in favour of such an approach.

Sampling from the Hypersphere On a k-dimensional hypersphere with radius r, caps are characterized by the polar angle θ, measured as the angle between rays from the center of the sphere to the pole and the base of the cap. The area of the spherical cap is given by [S, 2011] (assuming $\theta < \frac{\pi}{2}$)

$$A_k^{\text{cap}}(r,\theta) = \frac{1}{2}A_k(r)I_{\sin^2\theta}\left(\frac{k+1}{2},\frac{1}{2}\right) , \tag{A.8}$$

where

$$A_k(r) = \frac{2\pi^{k/2}}{\Gamma\left(\frac{k}{2}\right)}r^{k-1} . \tag{A.9}$$

denotes the area of the whole hypersphere, $\Gamma(y)$ the gamma function and $I_y(a,b)$ the incomplete beta function, both of which can be computed numerically. The factor $I_{\sin^2\theta}\left(\frac{n+1}{2},\frac{1}{2}\right)$ corresponds to the probability of receiving a point in the cap when sampling uniformly at random from the hypersphere.

Guarantees for passive sampling We now assume that we have trained a representation function $\hat{f}$ with Algorithm 1(i) and that we have trained a classifier $\hat{q}(x) = W\hat{f}(x)$ in the reconstructed latent space $\tilde{\mathcal{Z}}$. We want to derive error bounds for the representation function $\hat{f}$ in terms of its ability to recover the latent class structure. We assume that an $\hat{f}$ is an α-accurate minimizer[1] of the unsupervised training objective

$$\mathcal{L}_{un}(f) := \mathbb{E}_{(x,x^+,\{x_i^-\}_{i=1}^m)}\left[l\left(\{f(x)^T f(x^+) - \max_{1\le i\le m} f(x)^T f(x_i^-)\}_{i=1}^m\right)\right] , \tag{A.10}$$

which can be empirically estimated over a sample $\mathcal{D} = \{(x_j, x_j^+, x_{j1}^-, \ldots, x_{jm}^-)\}_{i=1}^n$ that contains n positive pairs and mn negative instances:

$$\hat{\mathcal{L}}_{un}(f) := \frac{1}{n}\sum_{j=1}^n l\left(\{f(x)^T f(x^+) - \max_{1\le i\le m} f(x)^T f(x_i^-)\}_{i=1}^m\right) . \tag{A.11}$$

[1] $z \in C$ implies $(g\circ f)(z) \in \tilde{C}$ with probability $1-\alpha$

We further define a supervised *margin loss* with respect to classifiers $q : \tilde{\mathcal{Z}} \to \mathbb{R}^{|\mathcal{C}|}$. For this, we first define a *margin function*

$$\gamma_q(f(x), c) := q_c(f(x)) - \max_{c' \neq c} q_{c'}(f(x)) \ . \tag{A.12}$$

With respect to a fixed margin $\gamma > 0$, we can define a loss function

$$\Phi_\gamma(v) := \min\left(1, \max\left(0, 1 - \frac{v}{\gamma}\right)\right) = \begin{cases} 1, & v \leq 0 \\ 1 - \frac{v}{\gamma}, & 0 \leq v \leq \gamma \\ 0, & \gamma \leq v \end{cases} , \tag{A.13}$$

and a margin loss

$$L_{class}(\mathcal{C}, q) := \mathbb{E}_{\substack{c \sim p_c \\ x \sim p_{data}(\cdot|c)}} \left[\Phi_\gamma\left(\gamma_q(f(x), c)\right)\right] \ . \tag{A.14}$$

We can empirically estimate L_{class} over $\mathcal{D}$ as

$$\hat{L}_{class}(\mathcal{C}, q, \mathcal{D}) := \frac{1}{n} \sum_{i=1}^{n} \Phi_\gamma(\gamma_q(f(x_i), c_i)) \ . \tag{A.15}$$

We can derive the following error bound:

Theorem A.3: Let $\gamma > 0$ (fixed) denote the margin in the true latent space and $\tilde{\gamma} \leq \gamma$ the margin in the reconstructed latent space. For any $\delta > 0$ we have with probability at least $1 - \delta$ that

$$L_{class}(f) \leq \frac{1}{1 - m\rho}\left(\hat{\mathcal{L}}_{un}(f) - m\rho\right) + \frac{1}{1 - m\rho}\left(4L_\alpha \mathcal{R}_{\mathcal{D}}(\mathcal{F}) + \sqrt{\frac{\log \frac{1}{\delta}}{2|\mathcal{D}|}}\right)$$

for all $f \in \mathcal{F}$. Here, $\mathcal{R}_{\mathcal{D}}(\mathcal{F})$ denotes the Rademacher complexity of the function class $\mathcal{F}$ and $L_\alpha \leq \frac{2}{\tilde{\gamma}}(1 - 2\alpha)$ and ρ the probability of sampling a false negative instance, i.e., the probability of sampling twice from the same class.
To compute the Rademacher complexity, we restrict f to the sample set $\mathcal{D}$ (with $|\mathcal{D}| =: n$)

$$f\big|_{\mathcal{D}} = \{\big(f(x_j), f(x_j^+), f(x_{1j}^-), \ldots, f(x_{mj}^-)\big)\}_{j=1}^n \subseteq \mathbb{R}^{3dmn} \ .$$

The Rademacher complexity is then given as

$$\mathcal{R}_{\mathcal{D}}(\mathcal{F}) = \mathbb{E}_{\sigma \sim \{\pm 1\}^{3dmn}}\left[\sup_{f \in \mathcal{F}} \langle \sigma,\ f\big|_{\mathcal{D}} \rangle\right] \ . \tag{A.16}$$

Remark A.4: The proof of Theorem A.3 is similar to [Saunshi et al., 2019, Theorem 4.1]. We give a bound with respect to an α-accurate representation function f, which approximates an optimal representation function f^* that recovers g^{-1} up to orthogonal linear transformation.
The proof of Theorem A.3 relies on two auxiliary lemmas, which we state first. Note that Φ_γ is $\frac{1}{\gamma}$-Lipschitz. This ensures the validity of the following standard bound for learning with noisy labels [Natarajan et al., 2013]) for the unsupervised contrastive training loss $\mathcal{L}_{un}$:

Lemma A.5 ([Natarajan et al., 2013]): For any fixed margin $\tilde{\gamma} > 0$ and a $\delta > 0$ we have with probability at least $1 - \delta$ over a ground truth set $\mathcal{D}$ for all $f \in \mathcal{F}$

$$\mathcal{L}_{un}(f) \leq \widehat{\mathcal{L}}_{un}(\mathcal{D}, f) + 4L_\alpha \mathcal{R}_{\mathcal{D}}(\mathcal{F}) + \sqrt{\frac{\log \frac{1}{\delta}}{2|\mathcal{D}|}} \ ,$$

where $\mathcal{R}_{\mathcal{D}}(\mathcal{F})$ denotes the Rademacher complexity of the function class $\mathcal{F}$ and $L_\alpha \leq \frac{2}{\tilde{\gamma}}(1 - 2\alpha)$.
We further need the following result, which relates the unsupervised contrastive loss $\mathcal{L}_{un}$ to the supervised loss L_{class}:

Lemma A.6: For any $f \in \mathcal{F}$ we have

$$L_{class}(\mathcal{C}, f) \leq \frac{1}{1 - m\rho}\left(\mathcal{L}_{un}(f) - m\rho\right) \ .$$

The lemma is a slight generalization of [Saunshi et al., 2019, Lemma 4.3].

$$\begin{aligned}
\mathcal{L}_{un}(f) &= \mathbb{E}_{c,c'\sim p_c,\ \substack{(x,x^+)\sim p_{pos}(\cdot|c)\\ x^-\sim p_{data}(\cdot|c')}} \left[\Phi_\gamma\left(f(x)^T f(x^+) - \max_{1\le i\le m} f(x)^T f(x^-)\right)\right] \\
&\ge \mathbb{E}_{\substack{c,c'\sim p_c\\ x\sim p_{data}(\cdot|c')}} \left[\Phi_\gamma\left(f(x)^T \mu_c - \max_{1\le i\le m} f(x)^T \mu_{c'}\right)\right] \\
&= (1-m\rho)\mathbb{E}_{\substack{c\neq c'\\ x\sim p_{data}(\cdot|c')}} \left[\Phi_\gamma\left(f(x)^T \mu_c - \max_{1\le i\le m} f(x)^T \mu_{c'}\right)\right] + m\rho \\
&= (1-m\rho)L_{class}(\mathcal{C}, f) + m\rho \ .
\end{aligned}$$

Then Theorem A.3 follows from combining Lemmas A.5 and A.6.

Comparison of passive and active sampling For our analysis of the active and passive sampling strategies, we pick two classes and focus on the problem of learning a separator between them. This reduces the problem to a binary classification task. Formally, let $|\mathcal{C}| = 2$ and

$$\begin{aligned}
C_+ &:= \{z|(z,1) \in \mathbb{S}^{k-1} \times \mathcal{C}\} \subseteq \mathcal{Z} \\
C_- &:= \{z|(z,-1) \in \mathbb{S}^{k-1} \times \mathcal{C}\} \subseteq \mathcal{Z} \ .
\end{aligned}$$

We assume that C_+, C_- are linearly separable with margin γ. Our goal is to to learn a classifier in $\tilde{\mathcal{Z}}$ that recovers the latent structure defined by $\{C_+, C_-\} \subseteq \mathcal{Z}$, i.e., we want to learn a classifier that separates $\{\tilde{C}_+, \tilde{C}_-\} \subseteq \tilde{\mathcal{Z}}$.

We want to compare the *passive* and *active* sampling approaches for positive instances. Note that the active sampling approach resembles classical active learning techniques for binary classification [Balcan et al., 2007], which allows us to utilize theoretical results from this literature. We show that the active selection approach reduces the number of samples that we need to add in each round (m_t), in comparison with the amount of samples needed, if positive instances are sampled passively. This indicates that representation functions can be trained more efficiently via active selection. Formally, we show the following result:

Theorem A.7: For any $\delta, \epsilon > 0$, we can recover the class structure up to error ϵ with probability $1-\delta$ with (1) sample complexity $m = O(\frac{d}{\epsilon})$ for random augmentations (Algorithm 1(i)) and (2) sample complexity $m = O\left(d^{3/2}\log(\frac{1}{\epsilon})\right)$ for active selection (Algorithm 1(ii), with rejection threshold $a_t = \frac{\pi}{2^{t-1}}$).

The proof follows results on active learning for binary classification [Balcan et al., 2007]. We outline the proof below. We will make use of the following standard result (see, e.g., [Anthony et al., 1999]):

Theorem A.8: Let H denote a set of functions from $\tilde{\mathcal{Z}}$ to ± 1 with finite VC dimension $V \ge 1$. Let D be a an arbitrary fixed distribution on $\tilde{\mathcal{Z}} \times \{\pm 1\}$. Then there exists a universal constant C, such that for any $\epsilon, \delta > 0$, if we draw a sample of size $N(\epsilon,\delta) = \frac{1}{\epsilon}(4V\log(\frac{1}{\epsilon}) + 2\log(\frac{2}{\delta}))$ from D, all hypotheses with error $\ge \epsilon$ are inconsistent with the data with probability $1-\delta$. (Thm. A.7)

(1) follows from Thm. A.8. For (2), we first note that the error of a classifier q can be measured with respect to w as

$$\mathrm{err}(w) = \frac{\arccos(w \cdot w^*)}{\pi} \ ,$$

where w^* denotes an optimal seperator for the data. With this, $\mathrm{err}(w) \le \epsilon$ implies $\|w - w^*\|_2 \le \epsilon\pi$. We want to show via induction that m_t samples are sufficient to obtain a classifier with $\mathrm{err}(w_t) \le 2^{-t}$ with probability $1 - \delta(1 - 1/(t+1))$. The case $t = 1$ follows again from Thm. A.8, i.e., with $m_1 = O(k + \log(1/\delta))$ we have $\mathrm{err}(w_1) \le \frac{1}{2}$ with probability $1 - \frac{\delta}{2}$. We assume that the claim is true for some t (induction hypothesis) and want to prove the claim for $t+1$. For an anchor point $x \sim p_{data}(\cdot|c)$ we can define the following two sets:

$$\begin{aligned}
S_1^t(x) &:= \{f(x^+) \in \tilde{\mathcal{Z}} : |w_t \cdot f(x^+)| \le a_t\} \\
S_2^t(x) &:= \{f(x^+) \in \tilde{\mathcal{Z}} : |w_t \cdot f(x^+)| > a_t\} \ .
\end{aligned}$$

In round t, we can write the error of the classifier q_t as

$$\mathrm{err}(w_t) = \mathrm{err}(w_t|S_1^t)P(S_1^t) + \mathrm{err}(w_t|S_2^t)P(S_2^t) \ ,$$

where $P(S)$ denotes the probability of sampling from S and

$$\mathrm{err}(w|S) := \mathrm{Prob}((w \cdot f(x))(w^* f(x)) < 0 | x \in S) \ .$$

Consider a classifier $\hat{w}$ that is consistent with $\mathcal{D}_t$. By the induction hypothesis both w_t and $\hat{w}$ have error at most 2^{-t}, i.e., $\mathrm{err}(\hat{w}) \leq 2^{-t}$ and $\mathrm{err}(w_t) \leq 2^{-t}$ with probability $1 - \delta(1 - 1/(t+1))$. This implies

$$\|w_t - w^*\|_2 \leq 2^{-t}\pi$$
$$\|\hat{w} - w^*\|_2 \leq 2^{-t}\pi \ .$$

Now let $\tilde{x} \in S_2$. Then

$$(w_t \cdot \tilde{x})(\hat{w} \cdot \tilde{x}) > 0$$
$$(w_t \cdot \tilde{x})(w^* \cdot \tilde{x}) > 0 \ ,$$

which implies $\mathrm{err}(\hat{w}|S_2) = 0$.

We can compute the probability $\mathrm{Prob}(\mathrm{S}_1^{\mathrm{t}})$ of sampling from the region S_1^t (close to the decision boundary) with respect to the acceptance threshold as

$$\mathrm{Prob}(S_1) \leq \frac{a_t\sqrt{k}}{2\pi} \ .$$

The proof follows from a geometric calculation and can be found in [Balcan et al., 2007, Lemma 4]. Inserting this above, we have

$$\mathrm{err}(\hat{w}) \leq 2^{-(t-1)}\sqrt{4\pi k} \cdot \mathrm{err}(\hat{w}|S_1) \ ,$$

which holds for all $\hat{w}$ consistent with $\mathcal{D}_t$. By construction, we add m_t samples (from S_1) to $\mathcal{D}_t$ in iteration t. By Thm. A.8 there exists a constant, such that with probability $1 - \frac{\delta}{t^2+t}$ we have

$$\mathrm{err}(\hat{w}|S_1) \leq \frac{1}{4\sqrt{4\pi k}}$$

for all $\hat{w}$ consistent with $\mathcal{D}_{t+1}$. This implies $\mathrm{err}(\hat{w}) \leq 2^{-(t+1)}$ for all $\hat{w}$ consistent with $\mathcal{D}_{t+1}$ and therefore the claim as $\mathrm{err}(w_{t+1}) \leq 2^{-(t+1)}$.

D. Experiments: Informative positives for latent classes

In the main text (sec. 4.2, Tab. 1), we present results for three sampling techniques.

Experimental setup. Throughout the experiments, we defined class-conditioned distributions with Gaussian noise of size 0.3. We sampled positive instances (*baseline*) and candidate positive instances (*active* and *double-active*) with perturbations of size 0.2. The experimental setup was a standard MLP, which was trained with learning rate 0.0001. The hyperparameters in the reported experimental results are listed in the main text. We investigate the following four scenarios, which employ different passive and active sampling strategies:

1) **Baseline:** Anchor points and positive instances are sampled from the baseline prior (resembling Alg. 1(i)).
2) **Active:** Anchor points are sampled from the baseline. Candidate positive instances are sampled via perturbation in the observation space until one is found whose image lies in the acceptance region, i.e., within some ϵ of the decision boundary of q_{t-1} in the reconstructed latent space. This resembles the *active selection* strategy (Alg. 1(ii)).
3) **Double-active:** Anchor points sampled with a bias towards the region near the decision boundary of q_{t-1} in the reconstructed latent space. Candidate positive instances are sampled via perturbation in the observation space until one is found whose image lies in the acceptance region. With the additional preference for sampling near the decision boundary, this can be seen as closer to the *adversarial augmentation* idea (Alg. 1(iii)).

E. Experiments: Targeted augmentation in latent space

Experimental setup. Tab. 2 in sec. 5.2 gives experimental results for the second experimental setting. Again, our experimental setup was a standard MLP, which was trained with learning rate 0.0001. We investigate the following sampling strategies:

1) **Baseline:** Equal-sized random perturbations of all latents.
2) **Info-active:** Small, targeted perturbations of informative latents and independent random sampling of all others.
3) **Active:** Small, targeted perturbations of informative latents and larger, targeted perturbations of all others.
4) **Class-preserving:** Assume access to a class-preserving transformation, which is used to augment the anchor to generate a positive sample from the same class. This is an idealized setting, which can be thought of as a supervised approach. It is included for comparison; unsupervised approaches are not expected to match its accuracy.

An important hyperparameter in the experiments is the *perturbation scale*, i.e., the size of the perturbations (Gaussian noise) in the "informative" and "noisy" (i.e., uninformative) dimensions. In the *baseline* approach, the perturbation scale is 0.3 across all latents. In the *Info-active* approach, we apply perturbations of 0.3 to the informative latent and 9.0 to the noisy latents. In

the *active* approach, we apply again perturbations of 0.3 to the informative latents, but only perturbations of 0.9 to the noisy latents.

Hyperparameter choice. We investigate the impact of the choice of the perturbation scale for the noisy latents on the knn accuracy. Fig. 3 shows results for perturbation scales of $0.3 - 9.0$ for different hyperparameters. We notice that even small differences in the perturbation scales between informative and noisy latents (i.e., small targeted perturbations of the informative latents) improve over the baseline.

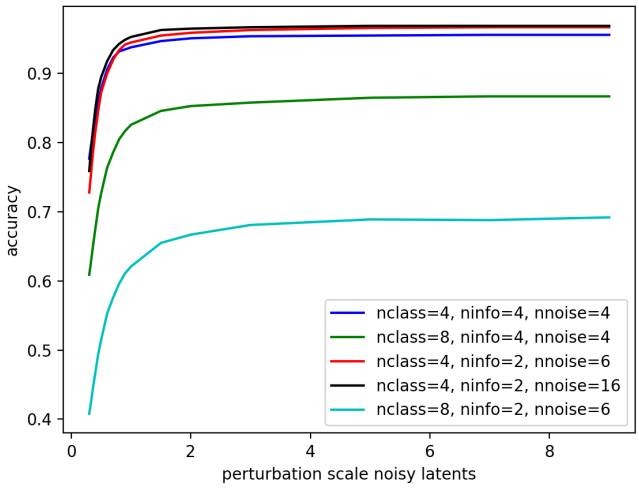

Fig. 3: Perturbation scale for "uninformative" or noisy latents.